Historical Contrastive Learning for Unsupervised Model Adaptation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Unsupervised domain adaptation aims to align a labeled source domain and an unlabeled target domain, but it requires to access the source data which often raises concerns in data privacy, data portability and data transmission efficiency. We study unsupervised model adaptation (UMA), an alternative setting that aims to adapt source-trained models towards target distributions without accessing source data. To this end, we design an innovative historical contrastive learning (HCL) technique that exploits historical source hypothesis to make up for the absence of source data in UMA. HCL addresses the UMA challenge from two perspectives. First, it introduces historical contrastive instance discrimination (HCID) that learns from target samples by contrasting their features which are generated by the currently adapted model and the historical models. With the source-trained and earlier-epoch models as the historical models, HCID encourages UMA to learn instance-discriminative target representations while preserving the source hypothesis. Second, it introduces historical contrastive category discrimination (HCCD) that pseudo-labels target samples to learn category-discriminative target representations. Instead of globally thresholding pseudo labels, HCCD re-weights pseudo labels according to their prediction consistency across the current and historical models. Extensive experiments show that HCL outperforms and complements state-of-the-art methods consistently across a variety of visual tasks ($e.g.$, segmentation, classification and detection) and setups ($e.g.$, close-set, open-set and partial adaptation).

1 Introduction

Deep neural networks (DNNs) [29, 62, 22] have achieved great success in various computer vision tasks [8, 43, 49, 48, 29, 62, 22] but often generalize poorly to new domains due to the inter-domain discrepancy [2]. Unsupervised domain adaptation (UDA) [67, 41, 65, 53, 55, 68, 66, 87, 86, 26, 60, 34, 27, 74] addresses the inter-domain discrepancy by aligning the source and target data distributions, but it requires to access the source-domain data which often raises concerns in data privacy, data portability, and data transmission efficiency.

In this work, we study unsupervised model adaptation (UMA), an alternative setting that aims to adapt source-trained models to fit target data distribution without accessing the source-domain data. Under the UMA setting, the only information carried forward is a portable source-trained model which is usually much smaller than the source-domain data and can be transmitted more efficiently [35, 32, 33, 61, 37] as illustrated in Table 1. Beyond that, the UMA setting also alleviates the concern of data privacy and intellectual property effectively. On the other hand, the absence of the labeled source-domain data makes domain adaptation much more challenging and susceptible to collapse.

To this end, we develop historical contrastive learning (HCL) that aims to make up for the absence of source data by adapting the source-trained model to fit target data distribution without forgetting

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Table 1: Source data have much larger sizes than source-trained models.

Storage size (MB)	Semantic segmentation		Object detection	Image classification
	GTA5	SYNTHIA	Cityscapes	VisDA17
Source dataset	62,873.6	22,323.2	12,697.6	7,884.8
Source-trained model	179.1	179.1	553.4	172.6

source hypothesis, as illustrated in Fig. 1. HCL addresses the UMA challenge from two perspectives. First, it introduces historical contrastive instance discrimination (HCID) that learns target samples by comparing their features generated by the current model (as queries) and those generated by historical models (as keys): a query is pulled close to its positive keys while pushed apart from its negative keys. HCID can thus be viewed as a new type of instance contrastive learning for the task of UMA with historical models, which learns instance-discriminative target representations without forgetting source-domain hypothesis. Second, it introduces historical contrastive category discrimination (HCCD) that pseudo-labels target samples for learning category-discriminative target representations. Instead of globally thresholding the predicted pseudo labels, HCCD re-weights the pseudo labels according to their consistency across the current and historical models.

The proposed HCL tackles UMA with three desirable features: 1) It introduces historical contrast and achieves UMA without forgetting source hypothesis; 2) The HCID works at instance level, which encourages to learn instance-discriminative target representations that are locally smooth and generalize well to new domains [81]; 3) The HCCD works at category level which encourages to learn category-discriminative target representation that is well aligned with the objective of down-stream tasks. We experimentally verify how these features facilitate UMA in Section 4.6 and appendix.

The contributions of this work can be summarized in three aspects. *First*, we investigate memory-based learning for unsupervised model adaptation that learns discriminative representations for unlabeled target data without forgetting source hypothesis. To the best of our knowledge, this is the first work that explores memory-based learning for the task of UMA. *Second*, we design historical contrastive learning which introduces historical contrastive instance discrimination and category discrimination, the latter is naturally aligned with the objective of UMA. *Third*, extensive experiments show that the proposed historical contrastive learning outperforms and complements state-of-the-art methods consistently across a variety of visual tasks and setups.

2 Related Works

Our work is closely related to several branches of research in unsupervised model adaptation, unsupervised domain adaptation, memory-based learning and contrastive learning.

Unsupervised model adaptation aims to adapt a source-trained model to fit target data distributions without accessing source-domain data. This problem has attracted increasing attention recently with a few pioneer studies each of which focuses on a specific visual task. For example, [35] freezes the classifier of source-trained model and performs information maximization on target data for classification model adaptation. [32] tackles classification model adaptation with a conditional GANs that generates training images with target-alike styles and source-alike semantics. [33] presents a self-entropy descent algorithm to improve model adaptation for object detection. [61] reduces the uncertainty of target predictions (by source-trained model) for segmentation model adaptation. [37] introduces data-free knowledge distillation to transfer source-domain knowledge for segmentation model adaptation. Despite the different designs for different tasks, the common motivation of these studies is to make up for the absence of source data in domain adaptation.

We tackle the absence of source data by a memory mechanism that encourages to memorize source hypothesis during model adaptation. Specifically, we design historical contrastive learning that learns target representations by contrasting historical and currently evolved models. To the best of our knowledge, this is the first work that explores memory mechanism for UMA. In addition, HCL is the first generic UMA method that works for a variety of visual tasks and adaptation setups.

Unsupervised domain adaptation is related to UMA but it requires to access labeled source data in training. Most existing work handles UDA from three typical approaches. The first exploits *adversarial training* to align source and target distributions in the feature, output or latent space

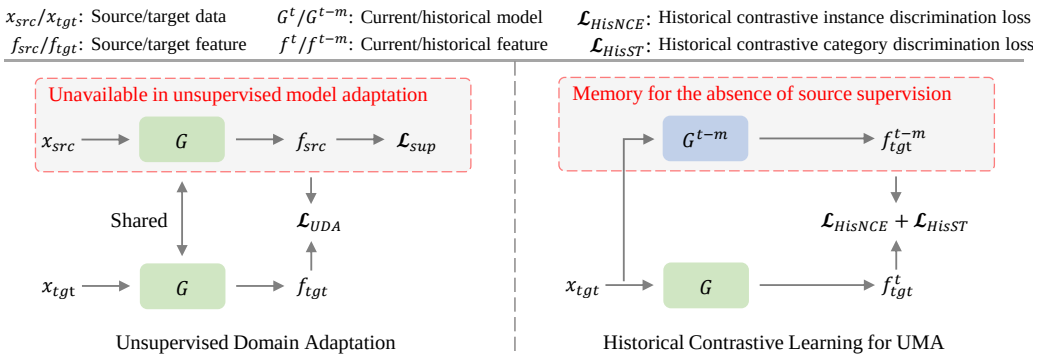

Figure 1: Illustration of unsupervised domain adaptation, unsupervised model adaptation and the proposed historical contrastive learning which exploits historical source hypothesis (or memorized knowledge) to make up for the absence of source supervision in the process of UMA. Here the historical source hypothesis could be the original source hypothesis G^0 ($i.e.$ t=m, trained using the labeled source data only) or the adapted source hypothesis G^{t-m} ($i.e.$ m<t, trained in the last m epoch). We use G^0 and G^{t-1} (the previous-epoch model) in our implementation.

[67, 41, 65, 14, 79, 53, 55, 68, 66]. The second employs *self-training* to generate pseudo labels to learn from unlabeled target data iteratively [87, 58, 83, 86]. The third leverages *image translation* to modify image styles to reduce domain gaps [26, 60, 34, 78, 27, 74].

Memory-based learning has been studied extensively. Memory networks [69] as one of early efforts explores to use external modules to store memory for supervised learning. Temporal ensemble [31], as well as a few following works [63, 12] extend the memory mechanism to semi-supervised learning. It employs historical hypothesis/models to regularize the current model and produces stable and competitive predictions. Mean Teacher [63] leverages moving-average models as the memory model to regularize the training, and similar idea was extended for UDA [16, 82, 5]. Mutual learning [80] has also been proposed for learning among multiple peer student models.

The aforementioned methods require labeled data in training. They do not work for UMA due to the absence of supervision from the labeled source data, by either collapsing in training or helping little in model adaptation performance. We design innovative historical contrastive learning to make up for the absence of the labeled source data, more details to be presented in the ensuing subsections.

Contrastive learning [70, 75, 21, 42, 85, 25, 44, 64, 9] learns discriminative representations from multiple views of the same instance. It works with certain dictionary look-up mechanism [21], where a given image x is augmented into two views, query and key, and the query token q should match its designated key k_+ over a set of negative keys k_- from other images. Existing work can be broadly classified into three categories based on dictionary creation strategies. The first creates a *memory bank* [70] to store all the keys in the previous epoch. The second builds an *end-to-end* dictionary [75, 64, 9] that generates keys using samples from the current mini-batch. The third employs a *momentum encoder* [21] that generates keys on-the-fly by a momentum-updated encoder. We design historical contrastive learning that generates keys by historical models.

3 Historical Contrastive Learning

This section presents the proposed historical contrastive learning that memorizes source hypothesis to make up for the absence of source data as illustrated in Fig. 2. The proposed HCL consists of two key designs. The first is historical contrastive instance discrimination which encourages to learn instance-discriminative target representations that generalize well to new domains [81]. The second is historical contrastive category discrimination that encourages to learn category-discriminative target representations which is well aligned with the objective of visual recognition tasks. The overall training process is provided in Algorithm.1, more details to be described in the ensuring subsections.

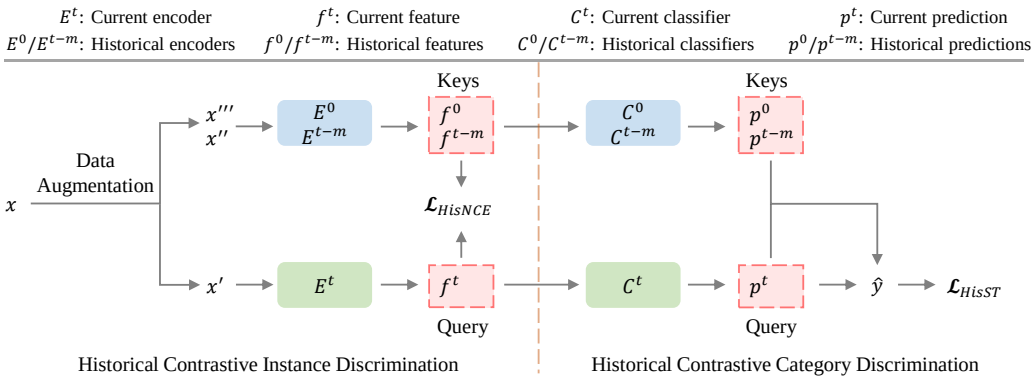

Figure 2: The proposed historical contrastive learning consists of two key designs including historical contrastive instance discrimination (HCID) and historical contrastive category discrimination (HCCD). HCID learns from target samples by contrasting their features generated by the current model (as queries) and historical models (as keys), which learns instance-discriminative target representations under the guidance of $\mathcal{L}_{\text{HisNCE}}$. HCCD pseudo-labels target samples to learn category-discriminative target representations under the guidance of $\mathcal{L}_{\text{HisST}}$, where the pseudo labels $\hat{y}$ are re-weighted adaptively according to the prediction consistency across the current and historical models.

3.1 Task Definition

We focus on the task of UMA. Given a source-trained model G and unlabeled target data X_{tgt}, the goal is to adapt G to perform well on X_{tgt}. The *baseline* is trained with the labeled source data:

$$\mathcal{L}_{sup}(X_{src}, Y_{src}; G) = l(G(X_{src}), Y_{src}), \tag{1}$$

where $l(\cdot)$ denotes an accuracy-related loss, $e.g.$, the standard cross-entropy loss. G is a visual task model that consists of an encoder E and a classifier C which classifies the encoded embeddings into pre-defined semantic categories, $i.e.$, $G(\cdot) = C(E(\cdot))$.

3.2 Preliminaries of Contrastive Learning

The idea of contrastive learning [20] can be considered as training an encoder (feature extractor) for a *dictionary look-up* task. Given a query q and a dictionary that consists of a number of keys $\{k_0, k_1, ..., k_N\}$, instance discriminative representations are learnt with an instance contrastive loss [20] (e.g. InfoNCE [44]), minimization of which will pull q close to its positive key k_+ and push it away from all other keys (considered negative for q):

$$\mathcal{L}_{\text{InfoNCE}} = \sum_{x_q \in X} -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{N} \exp(q \cdot k_i/\tau)} \tag{2}$$

where τ is a temperature parameter [70], q is the query representation by $q = E_{\text{q}}(x_q)$ where E_{q} is an query encoder network and x_q is a query sample (likewise in $k = E_{\text{k}}(x_k)$).

3.3 Historical Contrastive Instance Discrimination

The proposed HCID learns from unlabeled target samples via contrastive learning over their features generated from current and historical models: the positive pairs (features of the same sample) are pulled close while negative pairs (features of different samples) are push apart. It is a new type of contrastive learning for UMA, which preserves source-domain hypothesis by generating positive keys from historical models. HCID works at instance level and encourages to learn instance- discriminative target representations that are locally smooth [81] and generalize well to new domains.

HCID loss. Given an query sample x_q and a set of key samples $X_k = \{x_{k_0}, x_{k_1}, x_{k_2}, ..., x_{k_N}\}$, HCID employs current model E^t to encode the query $q^t = E^t(x_q)$, and historical encoders E^{t-m} to encode the keys $k_n^{t-m} = E^{t-m}(x_{k_n}), n = 1, \cdots, N$. With the encoded embeddings, HCID is

Algorithm 1 Historical Contrastive Learning for Unsupervised Model Adaptation.

Require: Target dataset X_{tgt}; source-trained model G^0
Ensure: Learnt parameters θ of task model G^{Max_epoch}
1: **for** $t = 1$ **to** Max_epoch **do**
2: **for** $iter = 1$ **to** Num_data **do**
3: ***Historical contrastive instance discrimination:*** Update G^t with $\{G^0, G^{t-1}\}$ as in Eq. 3
4: ***Historical contrastive category discrimination:*** Update G^t with $\{G^0, G^{t-1}\}$ as in Eq. 4
5: **end for**
6: **end for**
7: **return** G^{Max_epoch}

achieved via a historical contrastive loss $\mathcal{L}_{\text{HisNCE}}$, minimization of which pulls q close to its positive key k_+^{t-m} while pushing it apart from all other (negative) keys:

$$\mathcal{L}_{\text{HisNCE}} = \sum_{x_q \in X_{tgt}} -\log \frac{\exp(q^t \cdot k_+^{t-m}/\tau) r_+^{t-m}}{\sum_{i=0}^{N} \exp(q^t \cdot k_i^{t-m}/\tau) r_i^{t-m}} \tag{3}$$

where τ is a temperature parameter [70], r indicates the reliability of each key k_n^{t-m}, with which we reweight the similarity loss of each key to encourage to memorize well-learnt instead of poorly-learnt historical embeddings. In this work, we use the classification entropy to estimate the reliability of each key: $r = 1 - C(k)\log C(k)$. The positive key is the augmentation of the query sample, and all the rest are negative keys.

Remark 1 *Note $\mathcal{L}_{HisNCE}$ in Eq.3 has a similar form as the InfoNCE loss in Eq.2. InfoNCE can actually be viewed as a special case of HisNCE, where all the query and keys are encoded by the current model ($m = 0$) and the reliability is fixed ($r_i^{t-m} = 1, \forall i$). For HisNCE, we assign each key a reliability score to encourage to memorize the well-learnt historical embeddings only. It is also worth noting that Eq. 3 only shows historical contrast with one historical model for simplicity. In practice, we could employ multiple historical models to comprehensively distill (memorize) the well-learnt embeddings from them. In this work, we adopt two historical models including the source-trained model ($i.e.$, $m = t$) that provides the original source hypothesis for long-term memorization, and the model of of previous epoch ($i.e.$, $m = 1$) that provides evolved source hypothesis for short-term memorization. It is achieved by computing Eq.3 multiple times independently.*

3.4 Historical contrastive category discrimination

We design HCCD that generates pseudo labels conditioned on a historical consistency, $i.e.$, the prediction consistency across the current and historical models. HCCD can be viewed as a new type of self-training, where pseudo labels are weighted by the historical consistency instead of global thresholding. It works at category level and encourages to learn category-discriminative target representations that are aligned with the objective of visual recognition tasks in UMA.

Historical contrastive pseudo label generation. Given an unlabeled sample x, the current and historical models predict $p^t = G^t(x)$ (as the query) and $p^{t-m} = G^{t-m}(x)$ (as the keys). The pseudo label and the historical consistency of the sample are computed by:

$$\hat{y}^* = \arg\max_{\hat{y}} \sum_{k=1}^{K} \hat{y}_{(k)} \log p_{(k)}^t, \ s.t. \ \hat{y} \in \Delta^K,$$
$$h_{con} = 1 - \text{Sigmoid}(||p^t - p^{t-m}||_1)),$$

where p is a K-class probability vector and $\hat{y} = (\hat{y}^{(1)}, \hat{y}^{(2)}, ..., \hat{y}^{(C)})$ is the predicted category label.

HCCD loss. Given the unlabeled data x and its historical contrastive pseudo label $(\hat{y}, h_{con})$, HCCD performs self-training on target data x via a cross-entropy loss:

$$\mathcal{L}_{\text{HisST}} = -\sum_{x \in X_{tgt}} h_{con} \times \hat{y} \log p_x, \tag{4}$$

Table 2: Experiments on semantic segmentation task GTA5 → Cityscapes (SF: source free).

Method	SF	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg.	Terrain	Sky	PR	Rider	Car	Truck	Bus	Train	Motor	Bike	mIoU
Baseline [8]	✗	75.8	16.8	77.2	12.5	21.0	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36.0	36.6
CBST [87]	✗	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
AdaptSeg [65]	✗	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
AdvEnt [68]	✗	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
IDA [45]	✗	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0.0	30.2	35.8	46.3
CRST [86]	✗	91.0	55.4	80.0	33.7	21.4	37.3	32.9	24.5	85.0	34.1	80.8	57.7	24.6	84.1	27.8	30.1	26.9	26.0	42.3	47.1
UR [61]	✓	92.3	55.2	81.6	30.8	18.8	37.1	17.7	12.1	84.2	35.9	83.8	57.7	24.1	81.7	27.5	44.3	6.9	24.1	40.4	45.1
+HCL	✓	92.2	54.1	81.7	34.2	25.4	37.9	35.8	29.8	84.1	38.0	83.9	59.1	27.1	84.6	33.9	41.9	16.2	27.7	44.7	**49.1**
SFDA [37]	✓	91.7	52.7	82.2	28.7	20.3	36.5	30.6	23.6	81.7	35.6	84.8	59.5	22.6	83.4	29.6	32.4	11.8	23.8	39.6	45.8
+HCL	✓	92.3	54.5	82.6	33.1	26.2	38.9	37.9	31.7	83.5	38.1	84.4	60.9	30.0	84.5	32.6	41.2	14.2	26.4	43.2	**49.3**
HCID	✓	89.5	53	80.3	33.9	22.9	36.2	32.7	23.8	82.3	36.5	73.7	60.0	22.4	83.8	28.9	34.7	13.5	21.2	38.0	45.6
HCCD	✓	91.0	53.6	81.5	32.4	23.1	36.9	32.3	26.3	82.8	37.2	80.4	58.5	25.0	82.5	29.9	34.2	15.5	23.2	40.5	46.7
HCL	✓	92.0	55.0	80.4	33.5	24.6	37.1	35.1	28.8	83.0	37.6	82.3	59.4	27.6	83.6	32.3	36.6	14.1	28.7	43.0	**48.1**

Table 3: Experiments on semantic segmentation task SYNTHIA → Cityscapes (SF: source free).

Method	SF	Road	SW	Build	Wall*	Fence*	Pole*	TL	TS	Veg.	Sky	PR	Rider	Car	Bus	Motor	Bike	mIoU	mIoU*
Baseline [8]	✗	55.6	23.8	74.6	9.2	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5	38.6
CBST [87]	✗	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	78.3	60.6	28.3	81.6	23.5	18.8	39.8	42.6	48.9
AdaptSeg [65]	✗	84.3	42.7	77.5	-	-	-	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	-	46.7
AdvEnt [68]	✗	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2	48.0
IDA [45]	✗	84.3	37.7	79.5	5.3	0.4	24.9	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	41.7	48.9
CRST [86]	✗	67.7	32.2	73.9	10.7	1.6	37.4	22.2	31.2	80.8	80.5	60.8	29.1	82.8	25.0	19.4	45.3	43.8	50.1
UR [61]	✓	59.3	24.6	77.0	14.0	1.8	31.5	18.3	32.0	83.1	80.4	46.3	17.8	76.7	17.0	18.5	34.6	39.6	45.0
+HCL	✓	76.7	33.7	78.7	7.2	0.1	34.4	23.2	31.6	80.5	84.3	54.4	26.6	79.5	35.9	24.8	34.4	44.1	**51.1**
SFDA [37]	✓	67.8	31.9	77.1	8.3	1.1	35.9	21.2	26.7	79.8	79.4	58.8	27.3	80.4	25.3	19.5	37.4	42.4	48.7
+HCL	✓	78.2	35.3	79.6	7.3	0.2	37.7	21	30.9	80.4	83.3	59.8	29.4	79.2	34.2	24.5	38.9	45.0	**51.9**
HCL	✓	80.9	34.9	76.7	6.6	0.2	36.1	20.1	28.2	79.1	83.1	55.6	25.6	78.8	32.7	24.1	32.7	43.5	**50.2**

where h_{con} is the per-sample historical consistency and we use it to re-weight the self-training loss. If the predictions of a sample across the current and historical models are consistent, we consider it as a well-learnt sample and increase its influence in self-training. Otherwise, we consider it as pooly-learnt sample and decrease its influence in self-training.

Remark 2 *HCCD is an online algorithm which learns from all pseudo-labeled samples with different weights based on historical consistency, instead of learning from high-confidence samples only as identified by global thresholding [87, 86]. This new self-training strategy has unique features: 1) It is a threshold-free strategy that requires no hand-crafted thresholding which is computational expensive ($e.g.$, $3/4$ for pseudo-label generation while $1/4$ for learning from the generated pseudo-label); 2). It performs self-training over all samples instead of a small amount of selected ones which helps to mitigate class-imbalance and hard-easy-imbalance issues which could lead to over-fitting on selected samples and sub-optimal models.*

3.5 Theoretical Insights

The two designs in Historical Contrastive Learning (HCL) are inherently connected with some probabilistic models and convergent under certain conditions:

Proposition 1 *The historical contrastive instance discrimination (HCID) can be modelled as a maximum likelihood problem optimized via Expectation Maximization.*

Proposition 2 *The HCID is convergent under certain conditions.*

Proposition 3 *The historical contrastive category discrimination (HCCD) can be modelled as a classification maximum likelihood problem optimized via Classification Expectation Maximization.*

Proposition 4 *The HCCD is convergent under certain conditions.*

The proofs of **Propositions 1**, **2**, **3** and **4** are provided in the appendix.

4 Experiments

This section presents experiments including datasets, implementation details, evaluations of the proposed HCL in semantic segmentation, object detection and image classification tasks as well as the discussion of its desirable features.

Table 4: Experiments on object detection task Cityscapes → Foggy Cityscapes (SF: source free).

Method	SF	person	rider	car	truck	bus	train	mcycle	bicycle	mAP
Baseline [49]	✗	24.4	30.5	32.6	10.8	25.4	9.1	15.2	28.3	22.0
MAF [23]	✗	28.4	39.5	43.9	23.8	39.9	33.3	29.2	33.9	34.0
SCDA [84]	✗	33.5	38.0	48.5	26.5	39.0	23.3	28.0	33.6	33.8
DA [13]	✗	25.0	31.0	40.5	22.1	35.3	20.2	20.0	27.1	27.6
MLDA [71]	✗	33.2	44.2	44.8	28.2	41.8	28.7	30.5	36.5	36.0
DMA [30]	✗	30.8	40.5	44.3	27.2	38.4	34.5	28.4	32.2	34.6
CAFA [28]	✗	41.9	38.7	56.7	22.6	41.5	26.8	24.6	35.5	36.0
SWDA [54]	✗	36.2	35.3	43.5	30.0	29.9	42.3	32.6	24.5	34.3
CRDA [72]	✗	32.9	43.8	49.2	27.2	45.1	36.4	30.3	34.6	37.4
SFOD [33]	✓	25.5	44.5	40.7	33.2	22.2	28.4	34.1	39.0	33.5
+HCL	✓	39.3	46.7	48.6	32.9	46.2	38.2	33.9	36.9	**40.3**
HCL	✓	38.7	46.0	47.9	33.0	45.7	38.9	32.8	34.9	**39.7**

Table 5: Experiments on object detection task Cityscapes → BDD100k (SF: source free).

Method	SF	person	rider	car	truck	bus	mcycle	bicycle	mAP
Baseline [49]	✗	26.9	22.1	44.7	17.4	16.7	17.1	18.8	23.4
DA [13]	✗	29.4	26.5	44.6	14.3	16.8	15.8	20.6	24.0
SWDA [54]	✗	30.2	29.5	45.7	15.2	18.4	17.1	21.2	25.3
CRDA [72]	✗	31.4	31.3	46.3	19.5	18.9	17.3	23.8	26.9
SFOD [33]	✓	32.4	32.6	50.4	20.6	23.4	18.9	25.0	29.0
+HCL	✓	33.9	34.4	52.8	22.1	25.3	22.6	26.7	**31.1**
HCL	✓	32.7	33.2	52.0	21.3	25.6	21.5	26.0	**30.3**

4.1 Datasets

UMA for semantic segmentation is evaluated on two challenging tasks GTA5 [50] → Cityscapes [15] and SYNTHIA [51] → Cityscapes. GTA5 has $24,966$ synthetic images and shares 19 categories with Cityscapes. SYNTHIA contains $9,400$ synthetic images and shares 16 categories with Cityscapes. Cityscapes has 2975 and 500 real-world images for training and validation, respectively.

UMA for object detection is evaluated on tasks Cityscapes → Foggy Cityscapes [57] and Cityscapes → BDD100k [76]. Foggy Cityscapes is derived by applying simulated fog to the $2,975$ Cityscapes images. BDD100k has $70k$ training images and $10k$ validation images, and shares 7 categories with Cityscapes. We evaluate a subset of BDD100k ($i.e.$, *daytime*) as in [72, 54, 13] for fair comparisons.

UMA for image classification is evaluated on benchmarks VisDA17 [47] and Office-31 [52]. VisDA17 has $152,409$ synthetic images as the source domain and $55,400$ real images of 12 shared categories as the target domain. Office-31 has 4110 images of three sources including 2817 from **A**mazon, 795 from **W**ebcam and 498 from **D**SLR (with 31 shared categories). Following [86, 52, 59], the evaluation is on six adaptation tasks A→W, D→W, W→D, A→D, D→A, and W→A.

4.2 Implementation Details

Semantic segmentation: Following [65, 87], we employ DeepLab-V2 [8] with ResNet101 [22] as the segmentation model. We adopt SGD [4] with momentum 0.9, weight decay $1e-4$ and learning rate $2.5e-4$, where the learning rate is decayed by a polynomial annealing policy [8].

Object detection: Following [72, 54, 13], we adopt Faster R-CNN [49] with VGG-16 [62] as the detection model. We use SGD [4] with momentum 0.9 and weight decay $5e-4$. The learning rate is $1e-3$ for first $50k$ iterations and then decreased to $1e-4$ for another $20k$ iterations.

Image classification: Following [86, 52, 59], we adopt ResNet-101 and ResNet-50 [22] as the classification models for VisDA17 and Office-31, respectively. We use SGD [4] with momentum 0.9, weight decay $5e-4$, learning rate $1e-3$ and batch size 32.

We set the temperature τ at 0.07 as in [21]. All experiments are conducted in Pytorch with a single Tesla V100 GPU.

4.3 Unsupervised Domain Adaption for Semantic Segmentation

We evaluated the proposed HCL in UMA-based semantic segmentation tasks GTA5 → Cityscapes and SYNTHIA → Cityscapes. Tables 2 and 3 show experimental results in mean Intersection-over-Union (mIoU). We can see that HCL outperforms state-of-the-art UMA methods by large margins. In addition, HCL is complementary to existing UMA methods and incorporating it as denoted by

Table 6: Experiments on image classification benchmark VisDA17 (SF: source free).

Method	SF	Aero	Bike	Bus	Car	Horse	Knife	Motor	Person	Plant	Skateboard	Train	Truck	Mean
Baseline [22]	✗	60.9	21.6	50.9	67.6	65.8	6.3	82.2	23.2	57.3	30.6	84.6	8.0	46.6
MMD [38]	✗	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	85.8	20.7	61.1
DANN [17]	✗	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
ENT [18]	✗	80.3	75.5	75.8	48.3	77.9	27.3	69.7	40.2	46.5	46.6	79.3	16.0	57.0
MCD [55]	✗	87.0	60.9	83.7	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
CBST [87]	✗	87.2	78.8	56.5	55.4	85.1	79.2	83.8	77.7	82.8	88.8	69.0	72.0	76.4
CRST [86]	✗	88.0	79.2	61.0	60.0	87.5	81.4	86.3	78.8	85.6	86.6	73.9	68.8	78.1
3C-GAN [32]	✓	94.8	73.4	68.8	74.8	93.1	95.4	88.6	84.7	89.1	84.7	83.5	48.1	81.6
+HCL	✓	93.8	86.6	84.1	74.3	93.2	95.0	88.4	85.0	90.4	85.2	84.5	49.8	**84.2**
SHOT [35]	✓	93.7	86.4	78.7	50.7	91.0	93.5	79.0	78.3	89.2	85.4	87.9	51.1	80.4
+HCL	✓	94.3	87.0	82.6	70.6	92.0	93.2	87.0	80.6	89.6	86.8	84.6	58.7	**83.9**
HCL	✓	93.3	85.4	80.7	68.5	91.0	88.1	86.0	78.6	86.6	88.8	80.0	74.7	**83.5**

Table 7: Experiments on image classification benchmark Office-31 (SF: source free).

Method	SF	A→W	D→W	W→D	A→D	D→A	W→A	Mean
Baseline [22]	✗	68.4	96.7	99.3	68.9	62.5	60.7	76.1
DAN [38]	✗	80.5	97.1	99.6	78.6	63.6	62.8	80.4
RTN [39]	✗	84.5	96.8	99.4	77.5	66.2	64.8	81.6
DANN [17]	✗	82.0	96.9	99.1	79.7	68.2	67.4	82.2
ADDA [67]	✗	86.2	96.2	98.4	77.8	69.5	68.9	82.9
JAN [40]	✗	85.4	97.4	99.8	84.7	68.6	70.0	84.3
CBST [87]	✗	87.8	98.5	100	86.5	71.2	70.9	85.8
CRST [86]	✗	89.4	98.9	100	88.7	72.6	70.9	86.8
3C-GAN [32]	✓	93.7	98.5	99.8	92.7	75.3	77.8	89.6
+HCL	✓	93.4	99.3	100.0	94.6	77.1	79.0	**90.6**
SHOT [35]	✓	91.2	98.3	99.9	90.6	72.5	71.4	87.3
+HCL	✓	92.8	99.0	100.0	94.4	76.1	78.3	**90.1**
HCL	✓	92.5	98.2	100.0	94.7	75.9	77.7	**89.8**

"+HCL" improves the existing UMA methods clearly and consistently. Furthermore, HCL even achieves competitive performance as compared with state-of-the-art UDA methods (labeled by ✗in the column SF) which require to access the labeled source data in training.

Ablation studies. We conduct ablation studies of the proposed HCL over the UMA-based semantic segmentation task GTA5 → Cityscapes. As the bottom of Table 2 shows, incorporating either HCID or HCCD (into the *Baseline*) outperforms the *Baseline* by large margins. In addition, HCID and HCCD offer orthogonal self-supervision signals where HCID focuses on instance-wise discrimination between queries and historical keys and HCCD focuses on category-wise discrimination among samples with different pseudo category labels. The two designs are thus complementary and the combination of them in HCL produces the best segmentation.

4.4 Unsupervised Domain Adaptation for Object Detection

We evaluated the proposed HCL over the UMA-based object detection tasks Cityscapes → Foggy Cityscapes and Cityscapes → BDD100k. Tables 4 and 5 show experimental results. We can observe that HCL outperforms state-of-the-art UMA method SFOD clearly. In addition, incorporating HCL into SFOD improves the detection consistently across both tasks. Similar to the semantic segmentation experiments, HCL achieves competitive performance as compared with state-of-the-art UDA methods (labeled by ✗in column SF) which require to access labeled source data in training.

4.5 Unsupervised Domain Adaptation for Image Classification

We evaluate the proposed HCL over the UMA-based image classificat tasks VisDA17 and Office-31. Tables 6 and 7 show experimental results. We can observe that HCL outperforms state-of-the-art UMA methods clearly. In addition, incorporating HCL into state-of-the-art UMA methods further improve the classification accuracy consistently in both tasks. Similar to the semantic segmentation and object detection experiments, HCL achieves competitive performance as compared with state-of-the-art UDA methods (labeled by ✗) which require to access labeled source data in training.

4.6 Discussion

Generalization across computer vision tasks: We study how HCL generalizes across computer vision tasks by evaluating it over three representative tasks on *semantic segmentation*, *object detection* and *image classification*. Experiments in Tables 2- 7 show that HCL achieves competitive performance consistently across all three visual tasks without any task-specific optimization or fine-tuning. To the best of our knowledge, HCL is the first generic UMA method that works for a variety of visual tasks.

Table 8: Experiments on image classification benchmark Office-Home under the setup of partial-set DA (domain adaptation) and open-set DA.

Partial-set DA	SF	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Mean
Baseline [22]	✗	46.3	67.5	75.9	59.1	59.9	62.7	58.2	41.8	74.9	67.4	48.2	74.2	61.3
IWAN [77]	✗	53.9	54.5	78.1	61.3	48.0	63.3	54.2	52.0	81.3	76.5	56.8	82.9	63.6
SAN [6]	✗	44.4	68.7	74.6	67.5	65.0	77.8	59.8	44.7	80.1	72.2	50.2	78.7	65.3
ETN [7]	✗	59.2	77.0	79.5	62.9	65.7	75.0	68.3	55.4	84.4	75.7	57.7	84.5	70.5
SAFN [73]	✗	58.9	76.3	81.4	70.4	73.0	77.8	72.4	55.3	80.4	75.8	60.4	79.9	71.8
SHOT [35]	✓	57.9	83.6	88.8	72.4	74.0	79.0	76.1	60.6	90.1	81.9	68.3	88.5	76.8
+HCL	✓	66.9	85.5	92.5	78.3	77.2	87.1	78.3	65.1	90.7	82.4	68.7	88.4	**80.1**
HCL	✓	65.6	85.2	92.7	77.3	76.2	87.2	78.2	66.0	89.1	81.5	68.4	87.3	**79.6**
Open-set DA	**SF**	**A→C**	**A→P**	**A→R**	**C→A**	**C→P**	**C→R**	**P→A**	**P→C**	**P→R**	**R→A**	**R→C**	**R→P**	**Mean**
Baseline [22]	✗	36.3	54.8	69.1	33.8	44.4	49.2	36.8	29.2	56.8	51. 4	35.1	62.3	46.6
ATI-λ [46]	✗	55.2	52.6	53.5	69.1	63.5	74.1	61.7	64.5	70.7	79.2	72.9	75.8	66.1
OSBP [56]	✗	56.7	51.5	49.2	67.5	65.5	74.0	62.5	64.8	69.3	80.6	74.7	71.5	65.7
OpenMax [3]	✗	56.5	52.9	53.7	69.1	64.8	74.5	64.1	64.0	71.2	80.3	73.0	76.9	66.7
STA [36]	✗	58.1	53.1	54.4	71.6	69.3	81.9	63.4	65.2	74.9	85.0	75.8	80.8	69.5
SHOT [35]	✓	62.5	77.8	83.9	60.9	73.4	79.4	64.7	58.7	83.1	69.1	62.0	82.1	71.5
+HCL	✓	64.2	78.3	83.0	61.1	72.2	79.6	65.5	59.3	80.6	80.1	72.0	82.8	**73.2**
HCL	✓	64.0	78.6	82.4	64.5	73.1	80.1	64.8	59.8	75.3	78.1	69.3	81.5	**72.6**

Complementarity studies: We study the complementarity of our proposed HCL by incorporating it into existing UMA methods. Experiments in Tables 2- 7 (the row highlighted by "+HCL") show that incorporating HCL boosts the existing UMA methods consistently across different visual tasks.

Generalization across learning setups: We study how HCL generalizes across learning setups by evaluating it over two adaptation setups, $i.e.$, *partial-set adaptation* and *open-set adaptation*. Experiments in Table 8 show that HCL achieves competitive performance consistently across both setups with little setup-specific optimization or fine-tuning.

Comparisons with memory-based learning: We compared HCL with existing memory-based learning [69, 31, 63, 12, 16, 5, 82, 19] in UMA performance. Most existing memory-based learning methods perform certain temporal regularization, where the regularizer is generated from either historical models[69, 31, 63, 12, 16, 5] or the current model [82, 19]. Experiments over the UMA task GTA→Cityscapes (in Appendix) show that existing memory-based learning does not help in UMA. The major reason is that the existing memory-based learning requires to access labeled source data in training and it tends to collapse without such supervision. Similar collapse phenomenon has been observed in negatives-free unsupervised representation learning[11]. HCL also performs memory-based learning but it works effectively for UMA, largely because it memorizes historical hypothesis via instance contrastive learning that compares the query with its positive and negative keys, where the negative keys prevent collapse by enforcing the negative pairs apart [21, 9].

Comparisons with contrastive learning: We compared HCL with existing instance contrastive learning in UMA performance. Most existing instance contrastive learning methods perform certain dictionary look-up, where the dictionary is created by using memory bank [70], mini-batch [9, 20, 44, 25, 75, 1, 24] and momentum [21, 10]. Experiments over the UMA task GTA→Cityscapes (in Appendix) show that the existing instance contrastive learning does not perform well in UMA. The major reason is that the existing instance contrastive learning was designed for unsupervised representation learning without considering source hypothesis preservation. HCL also performs contrastive learning but works effectively for UMA, largely because it learns discriminative representation while persevering source hypothesis which is essential to UMA.

Due to the space limit, we provide the visualization of HCL feature representations and qualitative segmentation examples (and comparisons with state-of-the-art methods) in the appendix.

5 Conclusion

In this work, we studied historical contrastive learning, an innovative UMA technique that exploits historical source hypothesis to make up for the absence of source data in UMA. We achieve historical contrastive learning by novel designs of historical contrastive instance discrimination and historical contrastive category discrimination which learn discriminative representations for target data while preserving source hypothesis simultaneously. Extensive experiments over a variety of visual tasks and learning setups show that HCL outperforms and complements state-of-the-art techniques consistently. Moving forward, we will explore memory-based learning in other transfer learning tasks.

References

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

[2] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

[3] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016.

[4] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[5] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019.

[6] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Partial transfer learning with selective adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2724–2732, 2018.

[7] Zhangjie Cao, Kaichao You, Mingsheng Long, Jianmin Wang, and Qiang Yang. Learning to transfer examples for partial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2985–2994, 2019.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[10] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

[12] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Semi-supervised deep learning with memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–283, 2018.

[13] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.

[14] Yuhua Chen, Wen Li, and Luc Van Gool. Road: Reality oriented adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7901, 2018.

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[16] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[18] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005.

[19] Dayan Guan, Jiaxing Huang, Shijian Lu, and Aoran Xiao. Scale variance minimization for unsupervised domain adaptation in image segmentation. *Pattern Recognition*, 112:107764, 2021.

[20] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6668–6677, 2019.

[24] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

[25] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[26] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International Conference on Machine Learning*, pages 1989–1998, 2018.

[27] Weixiang Hong, Zhenzhen Wang, Ming Yang, and Junsong Yuan. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2018.

[28] Cheng-Chun Hsu, Yi-Hsuan Tsai, Yen-Yu Lin, and Ming-Hsuan Yang. Every pixel matters: Center-aware feature alignment for domain adaptive object detector. In *European Conference on Computer Vision*, pages 733–748. Springer, 2020.

[29] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[30] Taekyung Kim, Minki Jeong, Seunghyeon Kim, Seokeon Choi, and Changick Kim. Diversify and match: A domain adaptive representation learning paradigm for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12456–12465, 2019.

[31] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[32] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020.

[33] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. *arXiv preprint arXiv:2012.05400*, 2020.

[34] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019.

[35] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

[36] Hong Liu, Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Separate to adapt: Open set domain adaptation via progressive separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2927–2936, 2019.

[37] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. *arXiv preprint arXiv:2103.16372*, 2021.

[38] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015.

[39] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016.

[40] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

[41] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[42] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[43] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[44] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[45] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. *arXiv preprint arXiv:2004.07703*, 2020.

[46] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 754–763, 2017.

[47] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2021–2026, 2018.

[48] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[49] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[50] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016.

[51] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[52] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[53] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. *arXiv preprint arXiv:1711.01575*, 2017.

[54] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019.

[55] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723–3732, 2018.

[56] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–168, 2018.

[57] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018.

[58] Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, and Jose M Alvarez. Effective use of synthetic data for urban scene semantic segmentation. In *European Conference on Computer Vision*, pages 86–103. Springer, 2018.

[59] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2018.

[60] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018.

[61] Prabhu Teja Sivaprasad and Francois Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, number CONF. IEEE, 2021.

[62] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[63] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

[64] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[65] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018.

[66] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1456–1465, 2019.

[67] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

[68] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019.

[69] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[70] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[71] Rongchang Xie, Fei Yu, Jiachao Wang, Yizhou Wang, and Li Zhang. Multi-level domain adaptive learning for cross-domain detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[72] Chang-Dong Xu, Xing-Ran Zhao, Xin Jin, and Xiu-Shen Wei. Exploring categorical regularization for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11724–11733, 2020.

[73] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1426–1435, 2019.

[74] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.

[75] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6210–6219, 2019.

[76] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018.

[77] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8156–8164, 2018.

[78] Xiaobing Zhang, Haigang Gong, Xili Dai, Fan Yang, Nianbo Liu, and Ming Liu. Understanding pictograph with facial features: End-to-end sentence-level lip reading of chinese. *In AAAI*, pages 9211–9218, 2019.

[79] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2030, 2017.

[80] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[81] Nanxuan Zhao, Zhirong Wu, Rynson WH Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*, 2020.

[82] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. *arXiv preprint arXiv:1912.11164*, 2019.

[83] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–607, 2019.

[84] Xinge Zhu, Jiangmiao Pang, Ceyuan Yang, Jianping Shi, and Dahua Lin. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2019.

[85] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019.

[86] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5982–5991, 2019.

[87] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 289–305, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Please refer to the experiments Section for more details, where we experimentally verify the mentioned claims.
   (b) Did you describe the limitations of your work? [Yes] Please refer to experiment Section for more details, where we mentioned that our HCL is generic and can work for a variety of visual tasks and learning setups. In another word, the proposed method did not include any task-specific and setup-specific designs, where such extra task-specific or setup-specific designs can improve the performance for a specific task or setup greatly. We leave them for future works.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Following please find them. This work could mitigate ethical concerns of privacy violation for using images of source-domain. Current unsupervised domain adaptation works require using source-domain data in training which could raise concerns in data privacy. This work proposes HCL for unsupervised model adaptation, which aims to adapt source-trained models towards target distributions without accessing source data. Thus, it could relieve the reliance on source-domain data in training and so the risk of privacy violation. On the other hand, the UMA techniques (including ours and previous works) are still at the very early research stage. They could serve as certain assistant tools instead of any hard control in highly autonomous system ($e.g.$, autonomous-driving) which could lead to harmful consequences.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Please refer to the theoretic proofs in supplemental materials.

(b) Did you include complete proofs of all theoretical results? [Yes] Please refer to the theoretic proofs in supplemental materials.

3. If you ran experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] The code involves Intellectual Property issues. We plan to make the code public by submitting it during camera-ready submission, if this publication is accepted and the patent of this technique is filed. It is expected that this technique will be filed as a patent before the camera-ready DDL. The data is publicly accessible as mentioned in the paper.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please refer the Section 4 Experiments.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] In the manuscript, we follow the previous works to conduct the experiments for a consistent comparison, where error bars are not included in these previous works.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please refer to the implementation details in experiments Section.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes]

   (b) Did you mention the license of the assets? [Yes]

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] We did not use any new assets.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] For target-domain data, we used public datasets built by previous works, where all these datasets have their corresponding licenses. For source-domain data, we used synthetic datasets, which do not involves this issue.

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] For target-domain data, we used public datasets built by previous works, where all these datasets have their corresponding licenses. For source-domain data, we used synthetic datasets, which do not involves this issue.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use them.

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not use them.

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not use them.